Diphone-Based Concatenative Speech Synthesis System for Mongolian

Munkhtuya Davaatsagaan, and Kuldip K. Paliwal, *Member, IAENG*

Abstract— **This paper describes the first Text-to-Speech (TTS) system for the Mongolian language, using the general speech synthesis architecture of Festival. The TTS is based on diphone concatenative synthesis, applying TD-PSOLA technique. The conversion process from input text into acoustic waveform is performed in a number of steps consisting of functional components. Procedures and functions for the steps and their components are discussed in detail. Finally, the quality of synthesised speech is assessed in terms of acceptability and intelligibility.**

***Index Terms*—Diphone concatenation, Speech synthesis.**

I. Introduction

Text-to-speech synthesis enables automatic conversion of a sequence of type-written words into their spoken form. This paper deals with text-to-speech synthesis of Mongolian language. A few attempts have been made in the past to cover different aspects of a possible TTS system for Mongolian language [2] [3]. However, no-one has succeeded in building a complete system providing high quality synthesised speech. We have worked on text-to-speech synthesis for a year. Here we describe the current version of our Mongolian TTS system. Our improvements in the future will be based on this system.

The synthesis task is performed here through the following two steps: analyzing text and producing speech. Each of these steps includes several modules, operating sequentially, as shown in Fig.1 [1]. At the first step, input text is normalized in the Text processing module. The tasks of this module cover sentence tokenization, non-standard words and homograph disambiguation. In the phonetic analysis module, letter-to-sound rules are used for finding pronunciations of the normalised words. Their intonation, which includes accent, boundaries, duration and F0 is produced in the Prosodic analysis module. At the second step, the synthesiser creates a speech waveform from the complete phonetic and prosodic description. The diphone concatenative synthesis is used to generate a waveform from a sequence of phones by selecting and concatenating units from a prerecorded database of diphones. Modifying the pitch and duration to meet the prosodic requirements is performed by TD-PSOLA technique [6].

At the end of the paper, results of evaluation of the system are given and discussed and some promising directions for future work are mentioned.

Manuscript received December 29, 2007.

M. Davaatsagaan is with the School of Engineering, Griffith University, Brisbane, QLD 4111 Australia (phone: 614-023-57732; fax: 617-384-78963; e-mail: m.davaatsagaan@ griffith.edu.au).

K. K. Paliwal is with the School of Engineering, Griffith University, Brisbane, QLD 4111 Australia (e-mail: k.paliwal@ griffith.edu.au).

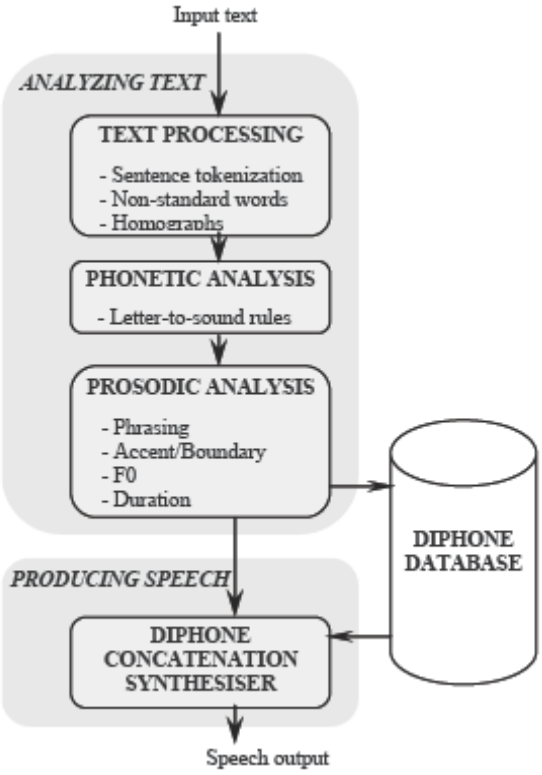

Figure 1: *General system architecture for Mongolian TTS system*

II. The Mongolian Language

Mongolian is the best-known member of the Mongolic language family, and the primary language of the residents in Mongolia. It is also spoken in some of the surrounding areas in provinces of China and Russia. The majority of speakers in Mongolia speak the Halh (or Khalkha) dialect. The Altaic theory proposes that the Mongolic family is a member of the larger Altaic family, which would also include the Turkic and Tungusic languages, and possibly Japanese and Korean [4]. Halh Mongolian is the national language of Mongolia. Around three million people speak Halh Mongolian throughout Mongolia. The Cyrillic alphabet is used for writing Mongolian. The alphabet was introduced in Mongolia by a government decision of 25 March 1941. It gradually replaced the old Mongolian script and was made the only official writing system in 1946 [5]. The script is the modern form of the Uigur Mongolian alphabet used for writing Mongolian since the thirteenth century.

III. Mongolian Tts Implementation

The task of speech synthesis is to map a text to a waveform. The speech synthesis system performs this process in two steps:

1. Analysing text: Converting the input text into a phonemic internal representation.

2. Producing speech: Converting the internal representation into a waveform.

The architecture of the system has a layered structure and each layer consists of functional components [1]. All required procedures and functions for the layers and their components will be defined in detail in the next sections.

A. Analysing Text

In this step, the input text is analysed and a phonemic internal representation is generated. Three modules (text processing, phonetic analysis and prosodic analysis) are used sequentially to carry out this step. These modules are described below.

A.1 Text processing

The text processing module performs sentence tokenization, handles the nonstandard words and carries out homograph disambiguation.

In Mongolian, like English, whitespace (space, tab, newline, and carriage return) and punctuation can be separated from the tokens in the text. Each identified token is mapped to words, standard or non-standard.

Non-standard words are tokens like numbers or abbreviations, which need to be expanded into sequences of Mongolian words before they are pronounced. These non-standard words are often very ambiguous. Dealing with non-standard words requires three steps: tokenization to separate out and identify potential non-standard words, classification to label them with a type from a predefined table, and expansion to convert each type into a string of standard words.

There are a few homographs in Mongolian. We need to disambiguate these homographs. Because knowledge of part-of-speech is sufficient to disambiguate some homographs, we store distinct pronunciations for these homographs labelled by part-of-speech, and then run a part-of-speech tagger to choose the pronunciation for a given homograph in context.

A.2 Phonetic analysis

The phonetic analysis module takes the normalized word strings from the text processing module and produces a pronunciation for each word. The pronunciation is provided not just as a list of phones, but also a syllabic structure and lexical stress. The method for finding the pronunciation of a word is either by a lexicon or by letter to sound rules. For Mongolian language, there exists a well-defined mapping from the orthography to the pronunciation, so lexicons for pronunciation are mostly redundant. Letter to sound rules do the whole job for almost all words in the Mongolian language.

For well defined languages like Mongolian, writing rules by hand is simpler than training. We built hand-written rule sets for this system. Hand written letter to sound rules are context dependent re-write rules which are applied in sequence mapping strings of letters to strings of phones.

A.3 Prosodic analysis

First, this module computes an abstract representation of the prosodic phrasing and pitch accent / boundaries of the text. Next, F0 values and duration are predicted from these prosodic structures.

An utterance has a prosodic phrase structure in a similar way to it having a syntactic phrase structure [7]. Simple rules based on punctuation are a very good predictor of prosodic phrase boundaries for Mongolian. We used the CART tree for predicting prosodic phrasing. This tree makes decisions based on distance from punctuation and whether the current word is a function word or content word.

Intonation is provided by a CART tree predicting ToBI (Tone and Break Indices) accents and an F0 contour generated from a model trained from natural speech. It used linear regression to assign target values to each syllable. For each syllable with a pitch accent or boundary tone, they predicted three target values, at the beginning, middle, and end of the syllable. They trained three separate linear regression models, one for each of the three positions in the syllable.

Phones vary quite a bit in duration. Some of the duration is inherent to the identity of the phone itself. However, phone duration is also affected by a wide variety of contextual factors [7]. A speech database was constructed to study the duration model of the Mongolian language. This database covers all the Mongolian phonemes and their most frequent contextual combinations. It contains words of various syllabic structures in various locations. For predicting duration, the system used a method which employs a tree to predict zscores that is the number of standard deviations from the mean. These zscores are used to calculate segmental duration following the formula below [8]:

$$\text{duration} = \text{mean} + (\text{zscore} * \text{standard deviation}) \quad (1)$$

The first process was to provide means and standard deviations for each phone in the Mongolian phoneset. The next process was to extract the features for predicting the durations. These features covered phonetic context, syllable, word position and type.

B. Producing Speech

At this step, speech waveform is created from a complete phonetic and prosodic description consisting of a list of phones associated with duration and a set of F0 targets. Diphone concatenative synthesis is used for creating waveforms in the system. The diphone concatenative synthesis model generates a waveform from a sequence of phones by selecting and concatenating diphones from a prerecorded database of diphones. A diphone is a phone-like unit going from roughly the middle of one phone to the middle of the following phone. In this section, we will describe building a diphone database and diphone concatenation.

B.1 Building Diphone database

First, we create an inventory of diphones for our system. Mongolian has 7 vowel and 32 consonant phonemes, so there are 39^2 =1521 hypothetically possible diphone combinations. Not all of these diphones actually occur. In addition, our system does not bother storing diphones if there is no possible coarticulation between the phones, such as across the silence between successive voiceless stops. Thus the system has only 944.

Next we recruited speakers who had some vocal talent. The current system has two voices, one male and one female. Then a text was created for the speakers to say, and record each diphone. In order to keep recording diphones as consistently as possible, each diphone was recorded to enclose in a carrier phrase. It was an unaccented nonsense word, pronounced with a steady intonation. By putting the diphone in the middle of other phones, we keep utterance-final lengthening or initial phone effects from making any diphone louder or quieter than the others. We used different carrier phrases for consonant-vowel, vowel-consonant, phone-silence, and silence-phone sequences.

Speech signals were recorded by a close talking microphone using a sampling rate of 16 kHz and 16 bit linear A/D conversion. After recording the speech, we labelled and segmented the two phones that make up each diphone. This was done automatically. But it was not completely accurate at finding phone boundaries, and so automatic phone segmentation was hand-corrected.

Finally, pitch markers were manually set for voiced parts of the corresponding speech signal.

B.2 Diphone concatenation

Given two diphones, in order to concatenate them, if the waveforms of the two diphones' edges across the juncture are very different, a perceptible click will result. Thus we applied a windowing function to the edge of both diphones so that the samples at the juncture have low or zero amplitude. Furthermore, if both diphones are voiced, the two diphones are joined pitch-synchronously. This means that the pitch periods at the end of the first diphone must line up with the pitch periods at the beginning of the second diphone; otherwise the resulting single irregular pitch period at the juncture is perceptible as well.

Now, given our sequence of concatenated diphones, in order to modify the pitch and duration to meet our prosodic requirements, we used TD-PSOLA (Time-Domain Pitch-Synchronous Overlap-and-Add) for the process.

Given an epoch-labeled corpus, the intuition of TD-PSOLA is that we can modify the pitch and duration of a waveform by extracting a frame for each pitch period, windowing the frame with a Hanning window, and then recombining these frames in various ways by simply overlapping and adding the windows pitch period frames. For assigning a specific duration to a diphone, to lengthen a signal with TD-PSOLA, we simply inserted extra copies of some of the pitch-synchronous frames, essentially duplicating a piece of the signal. For changing the F0 value of a recorded diphone, to increase the F0, we extracted each pitch-synchronous frame from the original recorded diphone signal, placed the frames closer together, with the amount of overlap determined by the desired period and hence frequency, and then added up the overlapping signals to produce the final signal. However, note that by moving all the frames closer together, we made the signal shorter in time. Thus to change the pitch while holding the duration constant, we added in duplicate frames.

IV. Evaluation

Some results of the performance assessment of the Mongolia TTS are given. The adequacy of the system was tested in two ways: in terms of acceptability and of intelligibility. The synthesis output was directed to a Sound Blaster audio card. The experiment was performed with 10 subjects aged between 32 and 47 years, half of them female. They were asked to judge 100-200 words, phrases, sentences, and real texts all harvested from internet sources such as newspapers, literature magazines and publications. In our first experiment, intelligibility of synthesised speech was evaluated on two levels: word level and sentence level. Subjects, participating in the test were asked to write down everything they heard. Fig. 2 gives the percentage of correctly understood words and sentences, with word intelligibility rate being close to 86 %. In our second experiment, degree of acceptability of the synthesised speech was assessed, again on word and sentence level.

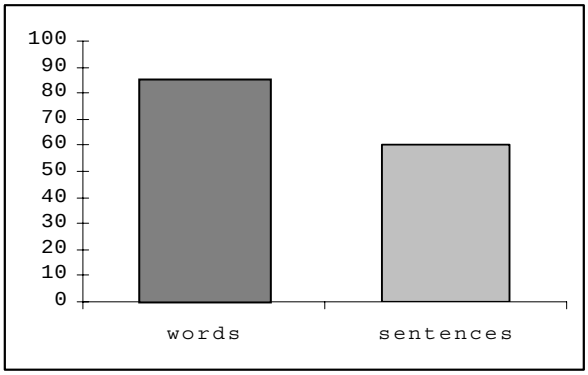

Figure 2: *Intelligibility test. Percentage of correctly understood words and sentences.*

Subjects were asked a few questions about naturalness, speed and sound quality and asked to mark how well the voice performs. The results are shown in Fig. 3, 4 and 5.

Regarding the question whether the voice is nice to listen to or not, 37% considered the voice natural, 40% thought that the naturalness of the voice was acceptable and 23 % considered the voice unnatural.

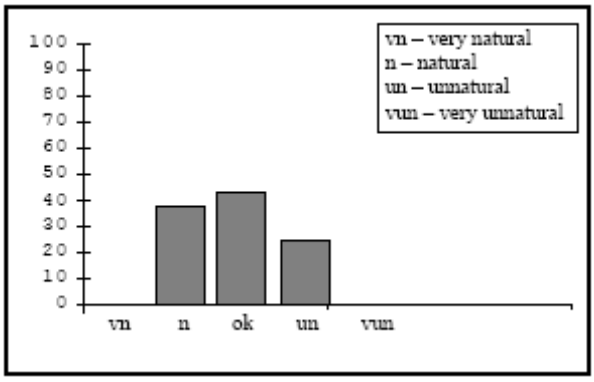

Figure 3: *Naturalness of the voice*

The speed of a system is a major concern, if the system speaks too fast or too slow this may have a negative effect on the concentration of the subjects. 76% of the listeners considered that the synthesised voice speed was normal rate.

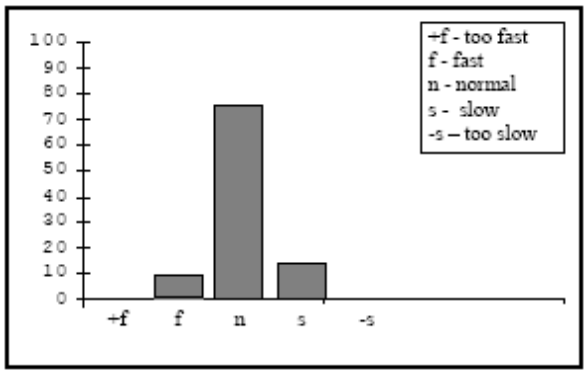

Figure 4: *Speed of the speech*

The question for this part was ”Did you consider the synthesised voice has good sound quality?”. 30% considered the voice has good quality; 60% thought the sound quality of the voice was neither bad nor good and the remaining 10% considered that the sound quality of the system bad.

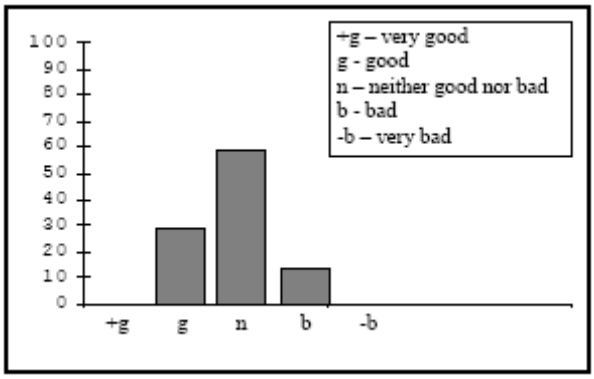

Figure 5: *Sound quality of the voice*

V. Conclusion

The described speech synthesis system will be the first complete TTS system for the Mongolian language. The earliest results are emerging. The synthetic speech produced by the current version of the system is intelligible, but utterances sometimes suffer from a lack of naturalness and fluency. Improvement of intelligibility and naturalness depends in particular on proper lexical stress assignment and a more sophisticated generation of prosodic features. The first attempts at developing a diphone-based synthesis system for the Mongolian language are promising, so that further work on improving individual parts of the system is encouraged. At this stage, we are working on research to improve on the simple models to create more sophisticated trainable models.

VI. Acknowledgements

The authors wish to thank the speakers for recording their voices and the subjects for the assessing process.

References

[1] B.Sukhbaatar, D.Munkhtuya, “A New Approach For The Mongolian Text-To-Speech Conversion” Proceedings of the International Conference on Electronics, Information, and Communication (ICEIC 2006), 2006, Vol. 2, pp.303 – 306, Ulaanbaatar, Mongolia

[2] Otgonbayar, B., *Study of all necessary parameters and characteristics of Mongolian speech synthesis system and establishing functional model for Mongolian spoken language*, Ph.D. thesis, Mongolian University of Science and Technology, 1996.

[3] BatEnkh, O., Investigation on New Approach to Build Speech Synthesis System for Mongolian Language. Ph.D. thesis, Mongolian University of Science and Technology, 2001.

[4] Poppe, N.N., Introduction to Altaic Linguistics: overview, Gengo no Kagaku (Science of Language), Vol. 6, pp. 130- 86, 1975a.

[5] Damdinsuren, Ts., Mongol shine usgiin tovch durem., Ulaanbaatar, 1946.

[6] Kleijn B., Paliwal K. (Editors), Elsevier Science B.V, *Speech Coding and Synthesis*, Netherlands, 1998.

[7] Black A., Taylor P., “The Festival Speech Synthesis System”, University of Edinburgh, Edinburgh, 2000 Available at http://www.cstr.ed.ac.uk/projects/festival

[8] Black A., Lenzo K., “Build Synthetic Voices”, Carnegie Mellon University, Edinburgh, 1999-2003